*End of National Polls:* Once the data extracted, we would need to process the polls data. Some studies (source ?) revealed that the only polls that were truly relevant for presidential elections are the ones that are issued for the last 6 months of a presidential campaign: those are the ones we selected. We subsequently aggregated these polls in order to get a *per-month* poll indicator for both Republican and Democrat candidates.

The final dataframe consisted of **36** observations for **35** features.

*Data Cleaning + Feature Engineering*: In order to clean our data, we first imputed the missing values by interpolation. We then checked the types of the columns and the consistency of values.

A crucial issue we needed to deal with, as mentionned earlier, was that we only had a few observations for a lot of features. Therefore, we tried to engineer this features. We created two new dataframes:

* The first one would consist in computing, for every section of our features (GDP, polls, NAP ..) a weighted sum of increases rate: where the weights would be more important as we are getting closer to the election. The resulting dataframe would consist of 36 obesrvations and **9** features. We call it the difference dataset
* The second one would use the features from the difference dataset and combine them with ‘weighted average’ features. In the same fashion as above, we would aggregate the features from a same indicator in a weighted sum, with weights getting more important as we are getting closer to election year. The resulting dataframe would consist of 36 obesrvations and **15** features.This is the sum/difference dataset.